Journal of Environmental Management 183 (2016) 98–109

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Review

A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods

O. Sarobidy Rakotonarivo [a, b, *], Marije Schaafsma [c, d], Neal Hockley [a]

[a] School of Environment, Natural Resource and Geography (SENRGY), Bangor University, LL57 2UW Bangor, Gwynedd, Wales, UK
[b] Food and Resource Economics, University of Copenhagen, Rolighedsvej 25, 1958 Frederiksberg C, Denmark
[c] Geography and Environment, University of Southampton, University Road, Southampton SO17 1BJ, UK
[d] Centre for Biological Sciences, University of Southampton, University Road, Southampton SO17 1BJ, UK

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
18 July 2016
Accepted 10 August 2016
Available online 27 August 2016

Keywords:
Discrete choice experiment
Validity
Reliability
Systematic review
Non-market environmental goods

ABSTRACT

While discrete choice experiments (DCEs) are increasingly used in the field of environmental valuation, they remain controversial because of their hypothetical nature and the contested reliability and validity of their results. We systematically reviewed evidence on the validity and reliability of environmental DCEs from the past thirteen years (Jan 2003–February 2016). 107 articles met our inclusion criteria. These studies provide limited and mixed evidence of the reliability and validity of DCE. Valuation results were susceptible to small changes in survey design in 45% of outcomes reporting reliability measures. DCE results were generally consistent with those of other stated preference techniques (convergent validity), but hypothetical bias was common. Evidence supporting theoretical validity (consistency with assumptions of rational choice theory) was limited. In content validity tests, 2–90% of respondents protested against a feature of the survey, and a considerable proportion found DCEs to be incomprehensible or inconsequential (17–40% and 10–62% respectively). DCE remains useful for non-market valuation, but its results should be used with caution. Given the sparse and inconclusive evidence base, we recommend that tests of reliability and validity are more routinely integrated into DCE studies and suggest how this might be achieved.

Contents

* Corresponding author. School of Environment, Natural Resource and Geography (SENRGY), Bangor University, LL57 2UW Bangor, Gwynedd, Wales, UK.
E-mail addresses: s.rakotonarivo@bangor.ac.uk (O.S. Rakotonarivo), m.schaafsma@soton.ac.uk (M. Schaafsma), n.hockley@bangor.ac.uk (N. Hockley).

http://dx.doi.org/10.1016/j.jenvman.2016.08.032

1. Introduction

It is frequently argued that improvements in environmental management require monetary valuation of environmental goods so that they are considered in decision-making (e.g. Jones-Walters and Mulder, 2009). Stated preference (SP) techniques offer an attractive valuation approach, particularly for environmental goods which are seldom traded in markets, and they have predictably become widely used for non-market valuation (Adamowicz, 2004). However, critics have long questioned their reliability and validity; that is whether they give consistent results across different survey designs that might be used to measure the same quantity and whether they measure what they are intended to (Bateman et al., 2002; Freeman, 2003). Their hypothetical nature is at the heart of the controversy: since respondents are asked to answer hypothetical questions, hypothetical bias may arise, i.e. respondents' expressed preferences may differ from their actual behaviour under real economic circumstances (Hausman, 2012).

The two most popular SP techniques are the contingent valuation method (CVM) and the discrete choice experiment (DCE) method (Freeman, 2003); the latter is the focus of this paper. CVM usually involves a single binary choice or open-ended question and was the dominant method for valuing non-market environmental goods in the 1990s. Latterly, DCEs have become widespread among environmental practitioners (Birol and Koundouri, 2008; Carson and Czajkowski, 2014). DCE originates in the market research and transport literature, and is rooted in Lancaster (1966) model of consumer choice, which proposed that the satisfaction that consumers derive from goods could be disaggregated into the good's various attributes. One of the main advantages of DCE over CVM is its ability to value the individual attributes characterizing a good or a policy, which may be more useful from a management perspective (Hanley et al., 2001). While DCE may potentially ameliorate some of the problems of CVM, it is likely to suffer from a number of similar limitations of CVM (Hanley et al., 2001) as well as new ones.

Issues with the reliability and validity of SP techniques (in particular CVM) have been widely acknowledged in textbooks, reviews and position papers (e.g. Mitchell and Carson, 1989; Bateman et al., 2002; Freeman, 2003; Carson and Hanemann, 2005; Carson and Groves, 2007). In particular, the design and analysis of DCE surveys have long been examined (e.g. Hanley et al., 1998; Louviere et al., 2000; Bennett and Blamey, 2001; Hensher et al., 2005; Louviere et al., 2011; Hess and Daily, 2014). Despite increasing efforts to tackle various reliability and validity aspects of DCE methods, DCE studies are still viewed with suspicion and debates are ongoing about various reliability and validity aspects even among SP practitioners (e.g. Hanley and Barbier, 2009; Carson and Groves, 2011; Hess and Daily, 2014; Lancsar and Swait, 2014). Nevertheless, DCEs remain widely used (e.g. Willis et al., 2003; Boatman et al., 2010; Christie et al., 2010). In order to reduce subjectivity, and given the controversies over the DCE method, it is vital that evidence on the reliability and validity of DCE studies is robustly synthesized so that those who might commission, conduct or rely upon their results in applied environmental settings comprehend its implications. Accordingly, this paper provides the first systematic review of empirical evidence from studies that have incorporated tests of the reliability or validity of the DCE method when valuing non-market environmental goods. This review also suggests areas for improvement and informs the development of contemporary guidelines in environmental DCE.[1] Applications of DCE in low-income and lower-middle-income countries[2] (LICs) may encounter further challenges to validity and reliability, as problems with low literacy rates, language barriers, difficulties in explaining hypothetical scenarios, and relatively low respondent exposure to surveys may be more prominent (Bennett and Birol, 2010; Christie et al., 2012). We have therefore specifically identified and highlighted evidence from, and implications for, DCEs conducted in LICs.

Systematic reviews have been developed in response to calls for a more rigorous and systematic approach to identifying and synthesising evidence that could inform policy (Haddaway and Pullin, 2014). Systematic reviews have the potential to enhance awareness of how much evidence is available in different parts of a field which can be useful for environmental management (e.g. Laurans et al., 2013). Unlike a conventional literature review, a systematic review follows a detailed, transparent, and reproducible search strategy, defined a priori (Pullin and Stewart, 2006), thereby aiming for completeness and objectivity in summarising the knowledge base. Systematic reviews have also been used to address methodological issues. However, in environmental management we are aware of only two systematic reviews that assessed methods: Petrokofsky et al. (2012) compared the accuracy and precision of methods for measuring carbon stocks, while Le Gentil and Mongruel (2015) assessed the methods and tools used to inform coastal zone management. While using systematic reviews to investigate the efficacy of research methods is still in its infancy, it may prove to be valuable for many methodological questions in environmental economics. We only know of two studies which used a systematic approach to review the application of SP methods in environmental valuation and these concentrated on the *usage* of CVM (Carson, 2011) and DCEs (Mahieu et al., 2014), rather than the reliability and validity of the methods. A secondary aim of this paper is therefore to consider the suitability of the systematic review approach for methodological questions in environmental valuation. In Section 2, we develop a conceptual framework for reliability and validity. Methods are presented in Section 3 and results are reported and discussed in Section 4, together with implications for researchers and decision-makers. We conclude in Section 5.

[1] Leading experts in the European Association of Environmental and Resource Economics (EAERE) are currently establishing such guidelines and standards for SP environmental valuation to promote broader acceptance of the method (see the session entitled "Emerging guidelines for stated preference methods in policy analysis" at the 21st Annual Conference).

[2] We used the World Bank's classification (http://data.worldbank.org/about/country-classifications/country-and-lending-groups accessed in August 2013). High income countries (HICs) refer to high income and upper-middle-income countries while LICs are low-income and lower-middle-income countries.

Table 1
Typology of validity and reliability testing in DCE studies.

Tests of			Methods
Reliability	Within-subject design		- Use of the test-retest approach at two different points in time - Use of deliberation or increased exposure to information
	Between-subject design (split sample)		- Small changes in the background scenario - Small changes of DCE attributes or levels - Use of different choice experiments designs.
Validity	External	Criterion	Comparison with actual (field) or simulated (laboratory) market experiments or non-hypothetical DCEs
		Convergent	Comparison with other methods such as hedonic pricing or contingent valuation
	Internal	Theoretical	Examination whether DCE responses conform to the standard axioms of rational choice theory: continuity (compensatory decision making as opposed to lexicographic or discontinuous preferences), transitivity, monotonicity, and stability (including order effects) Scope and embedding tests Use of qualitative techniques (e.g. verbal protocol or debriefing interviews or focus groups) to assess the above
		Content	Use of debriefing questions or qualitative techniques to assess respondent behaviour or perceptions: - Protest responses: trust towards the payment vehicle or belief in the credibility of the valuation scenario - Belief in the consequentiality of the survey - Respondent's stated or rated comprehension

2. Reliability and validity of the discrete choice experiment method: conceptual framework

The term “discrete choice experiment” is used throughout the review to avoid ambiguity, as suggested by Carson and Louviere (2011). The term “choice experiment” has different meanings in other disciplines such as biology and physics. To avoid confusion with the long-standing dichotomous CVM, we only cover DCE methods which involve more than a single choice set and allow analysts to estimate the marginal value of changing attributes as well as the total value of a good. Complete ranking techniques or other variants such as “best worst choice” or “pick any” techniques are often explicitly distinguished from DCE by SP researchers and are not covered by this systematic review, nor is “conjoint analysis” which originated from rating and rankings techniques that are generally inconsistent with economic demand theory (Louviere et al., 2010). Reliability refers to the degree of reproducibility of the results while validity refers to the degree to which the method is truly measuring what the researcher intended it to (Bateman et al., 2002; Freeman, 2003). It may not always be possible to clearly separate tests of reliability from validity tests because the two concepts are related; low reliability limits the overall validity of a test, and a lack of validity manifests itself in unreliable responses that vary with factors to which they should be robust (Davidshofer et al., 2005). The different types of validity tests are also not mutually exclusive but should be seen as focusing on different validity aspects. We have, however, attempted to distinguish them in the framework that follows. Table 1 summarizes the key concepts of reliability and validity testing.

2.1. Reliability

DCEs are reliable if they give consistent results across different surveys that might be used to measure the same quantity (Freeman, 2003). Studies testing for reliability usually survey the same individuals (within-subject design) or two independently drawn samples from the same population (between-subject or split-sample design). In the DCE literature, we identified five general ways to check for reliability: i) the test-retest approach using the same survey at two different points in time (e.g. Liebe et al., 2012; Schaafsma et al., 2014), ii) test of deliberation or greater exposure to information on DCE results (e.g. Robinson et al., 2008; Kenter et al., 2011), iii) test of framing effects or small changes in the background scenario (prior to choice sets) (e.g. Carlsson et al., 2010; Tonsor and Shupp, 2011), iv) test of small changes to DCE attributes or levels (e.g. Bateman et al., 2009; Solino et al., 2012), and v) comparisons of the results of different experimental design characteristics (e.g. Rolfe and Bennett, 2009; Baskaran et al., 2013). The first reliability check (i) is concerned with the temporal stability of stated values while the four others (ii to v) involve the simultaneous or subsequent use of two slightly different DCEs. The sensitivity of results to small changes in DCE survey instruments may be systematic and eventually predictable. Until then, we argue that these checks are important because decision-making often relies on the results of a single DCE survey. A systematic review of the outcomes of these tests therefore provides insights into the importance of methodological differences and how DCE surveys might usefully be improved.

2.2. Validity

Validity consists of i) external[3] validity (sometimes referred to as “concurrent validity” and including criterion and convergent validity) and ii) internal validity (theoretical and content validity). External validity tests involve comparisons with instruments other than a DCE survey while internal validity tests focus on the core assumptions of the DCE methods.

2.2.1. External validity testing

Criterion validity refers to the extent to which preferences elicited by the DCE method are related to another measure (a ‘criterion’) which is considered to be “true”, or at least closer to the theoretical construct of the investigation, such as data from real or simulated markets (Bateman et al., 2002). It is therefore directly concerned with hypothetical bias. However, for non-market environmental goods, the validity of market behaviour as a true measure of welfare might often be contested and for many environmental goods, no valid criterion measure can be observed. Therefore, some DCE researchers have used “real” or “non-hypothetical” DCE designs where respondents are presented with the same choices as in the hypothetical CE and then informed that one of the choices will be drawn randomly and will be binding, i.e. they will either have to pay or be paid the amount of money of the chosen alternative (e.g. Ready et al., 2010; Taylor et al., 2010).

Convergent validity refers to the correspondence between measures obtained by different methods (Freeman, 2003). In convergent validity testing, no method can be presumed superior to the other: two experiments that deliver the same estimates

[3] External validity in this review is different from the concept of external validity in the scientific literature generally, which refers to the extent to which the findings of a study can be legitimately transferred from one context to another (Brewer, 2000).

might just be equally invalid. DCE results can be compared with one of three alternatives: revealed preferences (e.g. travel cost models, production function approaches, hedonic pricing) (e.g. Scarpa et al., 2003); CVM or complete contingent ranking techniques (e.g. Caparros et al., 2008; Christie and Azevedo, 2009); or other valuation methods which may not be consistent with random utility theory such as multi-criteria analysis (e.g. Moran et al., 2007) or a simple attribute ranking exercise (e.g. Azevedo et al., 2009).

2.2.2. Internal validity testing

DCE results are said to be theoretically valid if respondents' choices do not deviate from the assumptions of standard rational choice theory (on which DCE methods are based), as defined by four axioms of utility maximisation (Mas-Colell et al., 1995). i) The "continuity axiom" refers to the use of compensatory decision-making rules i.e. attending to all the attribute levels across each of the alternatives and choosing the most preferred alternative within a choice task instead of using heuristics. Attribute non-attendance has also been referred to as discontinuous or lexicographic preferences (see Colombo et al., 2013 for a review in the environmental DCE literature). ii) Monotonic preferences require that, holding the levels of all other attributes equal, respondents should never prefer worse levels to better levels of an attribute (e.g. lower price in a WTP format should be preferred to a higher price). iii) The transitivity axiom requires that if a respondent prefers option A over option B and option B over option C, then he must prefer option A over option C. iv) The stability axiom[4] requires that when a respondent chooses an alternative A over an alternative B, he does not reverse his preference if presented with the same choice set later on. Stability testing also encompasses tests of order effects i.e. the influence of the order in which choice sets are presented to respondents (e.g. Day et al., 2012).

Other tests of theoretical validity concern sensitivity to scope. In DCE, sensitivity to scope broadly presumes that respondents should be willing to pay more for a large effect than for a subset of that effect (Carson and Czajkowski, 2014). Within-subject tests of sensitivity to scope assess whether a change in one or more attribute levels in a given alternative influence WTP significantly. Such within-subject tests may be judged to be weak; external scope tests which use a split sample design and compare WTP across samples from the same population are viewed as stronger tests (Rolfe and Wang, 2011). Scope tests are conceptually different to tests of monotonicity; failure to pass scope tests might not always indicate non-monotonicity: it may indicate satiation which is strictly compatible with the monotonicity axiom (Banerjee and Murphy, 2005). We included within subject and split sample scope tests.

Bateman et al. (2002, p305) refer to studies with high content validity as those in which the survey descriptions and questions are "conducive and sufficient to induce respondents to reveal valid stated values". We identified three measures of the content validity of the DCE method: i) protest attitudes, ii) comprehension of the DCE, and iii) perceptions of consequentiality. Measures of protest attitudes aim to identify respondents who object to some features of the survey or the valuation scenario and are distinguished from zero-bids. Protest attitudes often concern distrust towards the payment vehicle or beliefs regarding the credibility of the policy scenario (e.g. Meyerhoff and Liebe, 2009). Respondents' comprehension of the valuation exercise is either self-reported by respondents or rated by researchers (e.g. Barkmann et al., 2008). Measures of perceived consequentiality examine whether respondents care about the survey outcomes and view them as consequential: i.e. having real policy impact (e.g. Vossler et al., 2012).

Lack of theoretical and content validity can be identified in respondents' choices or self-reported by respondents in follow-up statements. The lack of validity has been measured by: i) the percentage of respondents showing violations of the utility axioms or perceiving a lack of content validity, ii) the effect on willingness-to-pay (WTP) estimates of, for example, removing the inconsistent choices from the analysis, or iii) entering an additional variable into the econometric specification that captures the lack of theoretical or content validity (e.g. Alemu et al., 2013). Qualitative methods can also be used to assess both theoretical and content validity of DCE. These include verbal protocols during the completion of the valuation task (Arana and Leon, 2009) or debriefing interviews after the DCE exercise (through focus groups or individual qualitative interviews) (e.g. Powe et al., 2005).

3. Methods

The systematic review process generally comprises five steps: the development of a protocol to guide the review, screening or inclusion criteria, quality appraisal, data extraction, and synthesis (Pullin and Stewart, 2006). As the primary objective of this review is to examine the evidence on the reliability and validity of the DCE method, we selected studies which met the inclusion criteria and whose survey design is judged sufficiently robust to answer our review question, but did not further appraise the quality[5] of the selected articles given the limited evidence base. We sent the review protocol to six DCE experts and practitioners, three of them reviewed it and provided valuable comments on the selection criteria and search strategy.

3.1. Systematic review protocol and search strategy

We used the conceptual framework developed in Section 2 to generate a set of search terms that were included in a search string formatted according to requirements for searching in the Web of Science (WoS) and EconLit databases. Following experts' recommendations, we used a set of 24 references (Supplement 1) as a 'test library' to check whether the search strings captured the expected studies, and, if not, what terms would have included them and how many other relevant studies using those new terms might add. We used an iterative checking process to validate the search terms and reduce the risk of missing relevant studies. The final search string employed (Fig. 1) was defined after 15 iterations and was judged to be sufficiently diverse to capture different phrasings of the reliability and validity of DCE. The search terms ensured a balance between the proportion of hits that are relevant (referred to in the systematic review literature as "specificity") whilst ensuring that all available literature was captured ("sensitivity"). We conducted the initial search between 20 July and 20 August 2013 by entering the search terms (Fig. 1) into two databases: i) the ISI Web of Science (WoS) (*https://webofknowledge.com/*), one of the world's largest databases of scientific papers and (ii) Econlit *(*http://www.aeaweb.org/econlit/icon.php*)*, the database of the American Economics Association which is a database of both peer reviewed literature and working papers in economics. The search was updated on 24–29 February 2016, using WoS only, since EconLit had returned

[4] Stability here is different from the temporal reliability defined in Section 2.1. In practice the difference is between stability within a survey (i.e. across different presentations at the same time) versus stability across identical presentations over time.

[5] Quality appraisal involves the scoring of each relevant study against a set of pre-established criteria or "quality hierarchy". These criteria often involve subjective judgements about the relative importance of different sources of bias (for more details, please see Pullin and Stewart, 2006).

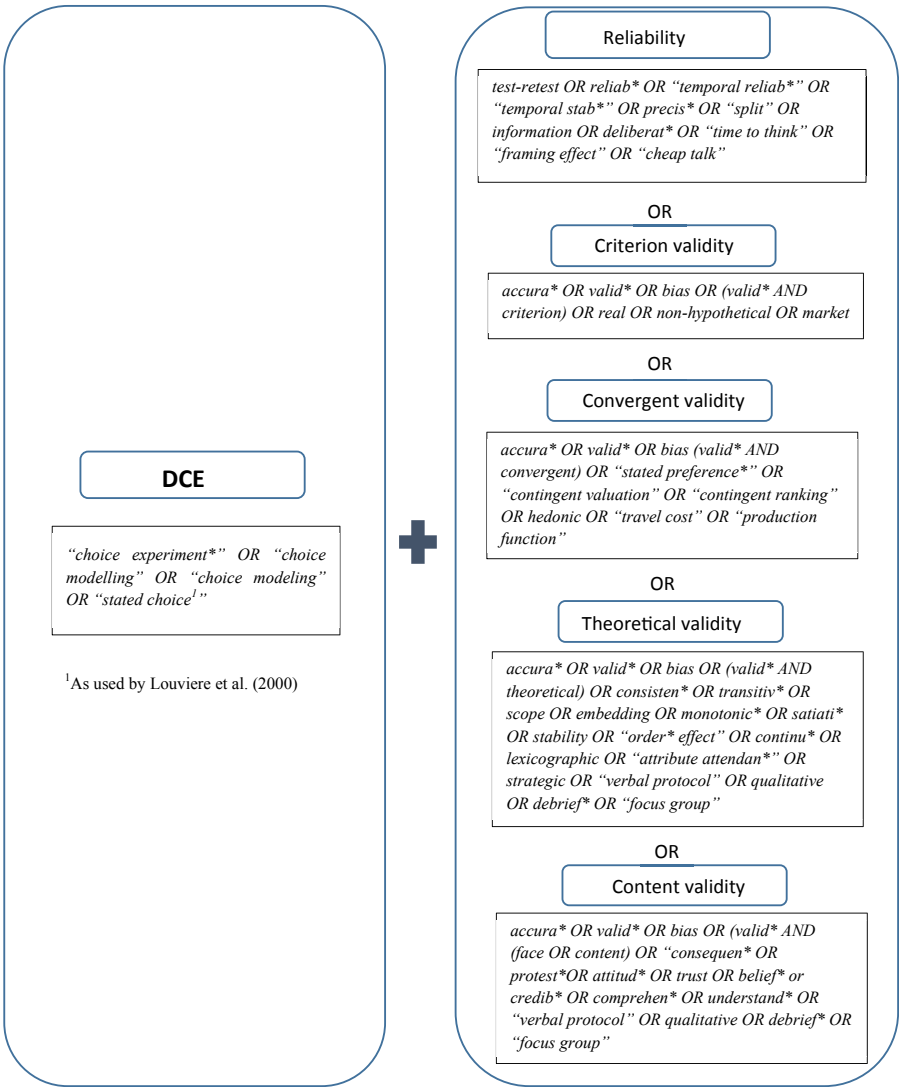


Fig. 1. Search strings (combination of sub-strings from DCE and different approaches to reliability and validity testing using Boolean operators).

only three additional includable articles in the initial search. After removing duplicates, articles were assessed against our inclusion criteria (see 3.2) first using titles and keywords, then abstracts, then full texts. At each stage any potentially includable studies were retained for the next stage. Included studies are described in the synthesis tables (Supplement 3), which report the type of validity and reliability checks, the good valued, the location, the sample design and sample size, the econometric methods used and the methods used to test for the equality of marginal willingness to pay (MWTP)/willingness to accept (MWTA) estimates.

3.2. Inclusion criteria, data extraction and synthesis

To be included in the review, studies had to satisfy the following criteria. They had to test for the validity or reliability of the DCE results, and must have been published between January 2003 and February 2016. The time span was restricted to capture the most recent studies as DCE and SP techniques have advanced over the years and are evolving fast. The object of valuation or type of good being valued was restricted to non-market environmental goods or non-market environmental attributes of market goods, including both use and non-use values. "Non-market" refers to goods that do not have an observable market price and are not sold or bought directly in the market (e.g. the regulation of water or air quality, or recreational and spiritual benefits – See Millenium Ecosystem Assessment, 2005). Non-market attributes of market goods include for instance the ecological component of certified coffee beans (e.g. Carlsson et al., 2010; Tonsor and Shupp, 2011), where organic production may be supposed to produce public goods as well as private benefits to the consumer. Only original DCE applications were included in the analysis, and benefit transfer studies, meta-analyses or discussion papers were excluded. Only papers in English were included.

To be included, qualitative studies must have explicitly reported results in a manner which allows an assessment of reliability/validity to be made. Studies which carried out focus groups or other qualitative methods simply to assist in drafting DCE surveys were excluded. Studies which only included robustness checks (Smith, 2007), which examine model fits or the robustness of results to different assumptions such as the treatment of unobserved heterogeneity or different model specifications (e.g. Campbell et al., 2011; Christie and Gibbons, 2011; Torres et al., 2011) were also excluded. Instead, we focused on the design and administration of DCE surveys, and on how respondents perceive and answer them, rather than on data analysis. Similarly, we excluded studies that only tested common prior expectations such as the relationship

between WTP estimates and income (Bateman et al., 2002). Such tests are routinely handled in data analyses and are ambiguous tests of validity.[6] We excluded respondents' self-reported certainty about their choices since low certainty may represent a real feature of respondents' preferences not a lack of validity. Likewise, we excluded comparisons of MWTP and MWTA estimates because the WTP-WTA disparity is not prima facie evidence of lack of reliability of the DCE method but may instead reflect underlying preferences consistent with Hicksian theory (Kim et al., 2015). Conversely, while comparing the effect of alternative survey administration modes on DCE results (e.g. Olsen, 2009) rightly qualifies as reliability testing, it is beyond of the scope of this systematic review which focused on survey design.

Different outcome elements were extracted from the included studies depending on the types of reliability or validity tests. Reliability, criterion and convergent validity testing often produce comparisons of attribute parameters (or utility coefficients), MWTP/MWTA or compensating surplus estimates between split samples. When comparing attribute parameters between two samples, we included outcomes which used the Swait and Louviere (1993) sequential testing procedure to account for differences in scale factors.[7] In logit models, the scale parameter (inversely related to the variance of the error term) is jointly estimated and hence confounded with the attribute parameters in the utility function (Louviere et al., 2000). Three tests for equality of MWTP/MWTA estimates were used in the reviewed studies; i) confidence intervals, ii) performing a simple *t*-test, and iii) using the complete combinatorial method (Poe et al., 2005). The first two tests can give biased outcomes if normality assumptions are violated: t-tests in particular might underestimate the level of significance of differences in WTP (*ibid*). Nevertheless, we included studies that used any of the three tests, but noted the approaches used by authors (Supplement 3). Studies are too heterogeneous to permit a quantitative meta-analysis. Instead, using the full synthesis tables (Supplement 3), we describe the state of evidence by highlighting the number of studies providing a yes or no answer to the questions of interest. We do not present effect sizes, which would be uninformative because both the context and the non-market environmental good being valued differed across studies.

4. Results and discussion

Searches in August 2013 returned 2350 articles from WoS and 2600 from Econlit. After removal of duplicates 2865 articles remained, and 995 of these were identified as potentially relevant from the title and keywords. 285 articles were retained after abstract-level screening, and 78 after initial full text assessment. The updated search in February 2016 resulted in 1104 articles, of which 59 articles were fully assessed. In total, 107 articles (29 were from the update) were included after the final stage of full text assessment, from which the outcomes of interest were extracted. The most common reasons for the exclusion of articles at this final stage included not valuing non-market environmental goods or non-market environmental attributes of market goods (e.g. Lusk and Schroeder, 2004; Hess et al., 2012), absence of a test of reliability or validity (e.g. Beharry-Borg et al., 2009; Bush et al., 2009), including only robustness checks (e.g. Campbell et al., 2011; Christie and Gibbons, 2011; Torres et al., 2011), discussions or theoretical articles (e.g. Carson and Groves, 2007; Carlsson, 2010), or not using a hypothetical DCE (e.g. Gracia et al., 2011; Michaud et al., 2013). Of the 107 studies retained, 12 articles (11%) were conducted in LIC and one is a working paper, the remainder were all in peer-reviewed journals. Supplement 2 indexes all 107 studies by their IDs; these studies are synthesized in supplement 3 and the studies excluded at the final stage of full text assessment along with the reasons for exclusion are reported in Supplement 4.

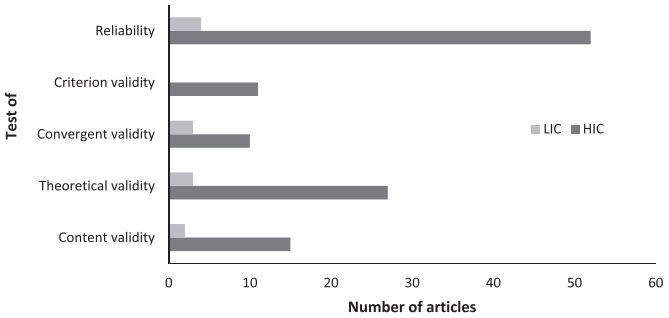


Fig. 2. Number of articles incorporating validity and reliability tests in low-income (LIC) and high-income (HIC) countries.

We found 56 and 65 articles incorporating reliability and validity tests respectively (14 tested both) (Fig. 2). Twenty-eight articles produced more than one outcome of reliability and/or validity tests, the total number of test outcomes was 173 (93 and 80 outcomes of reliability and validity tests respectively).

4.1. Reliability

Of the 87 outcomes of reliability tests (from 50 articles, of which only three were conducted in LICs), 39 (45%) found a significant difference between treatments: 20 (out of 50) for MWTP/MWTA and 19 (out of 37) for attribute parameters (Fig. 3). Six outcomes (from six articles, all but one in HICs) were neither comparisons of attribute parameters nor MWTP/MWTA estimates. Respondents' choices were not altered by deliberation in a HIC setting (S75), whereas the good valued became incommensurable with money following deliberation in a LIC context (S51). A "cheap talk script" significantly increased the percentage of respondents who chose the status quo option (S14 and S60). Likewise, different design characteristics (number of choice sets, alternatives, attributes, levels and the range of levels) reflecting different levels of complexity significantly affected choice outcomes (S14 and S66). Only three outcomes were derived from a WTA survey (S10, S50 and S39) and 67 outcomes (72%) from a between-subject design. This mixed evidence on the reliability of DCE is not unexpected, since survey research has long demonstrated that small changes in the design or wording can significantly affect outcomes (Schuman and Presser, 1981). It should also be remembered that statistically significant results may not be economically significant, and vice versa. Nevertheless, if two similar designs (each of which might be considered good practice) yield different results, decision-makers must apply appropriate caution in relying on the results of any single DCE study.

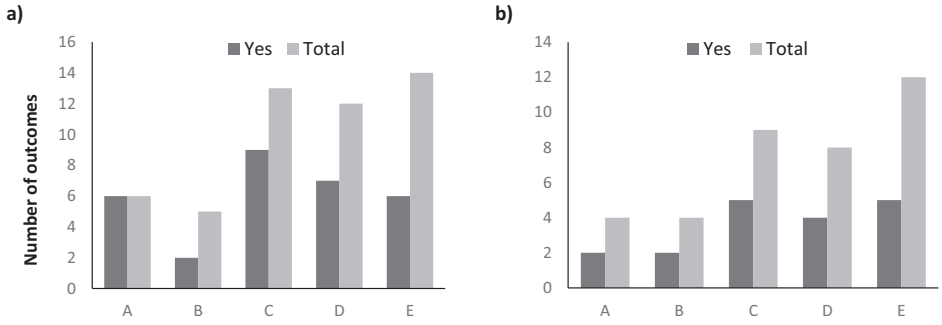
4.2. Validity

4.2.1. Do DCEs predict behaviour in real transactions?

Eleven articles used some criterion validity testing (S4, S14,

[6] We distinguish such tests from those described in Section 2.2.2, which concern assumptions on which the DCE method is based.

[7] We note that in addition to the Swait-Louviere sequential procedure, there are also less common methods used by other fields (transport and health economics) to control for scale differences such as the procedure proposed by Ben-Akiva and Morikawa (1990), in which observations from separate (groups of) choice tasks are used simultaneously to maximize a joint likelihood function; and the Bradley and Daly (1994) one-step estimation approach of Ben-Akiva and Morikawa, which can be implemented using a nested logit (the logit-based scaling approach).



A: Test-retest: Are DCE results stable over time?
B: Are DCE results robust to deliberation?
C: Are DCE results invariant to small changes in the background scenario prior to choice sets?
D: Do small changes of DCE attributes or levels have no effect on the DCE results?
E: Are DCE results unsusceptible to different design characteristics?

Fig. 3. a) Outcomes comparing MWTP/MWTA estimates, b) Outcomes comparing equality of attribute parameters (at 5% significance level) – Note: Yes responses to questions A – E indicate evidence consistent with the reliability of DCE i.e. attribute parameters or MWTP/MWTA estimates were not significantly different at the 5% level.

S17, S47, S48, S60, S71, S76, S94, S102, S106) producing 13 outcomes (Fig. 4). They were all conducted in HICs and 11 outcomes used non-hypothetical DCE as the criterion. None of these 11 outcomes supported the criterion validity of DCE: hypothetical bias varied from 50% to 100%. However, three of these 11 studies (S14, S60, and S71) also used a "cheap talk script" to "mitigate" hypothetical bias but only the first two succeeded (i.e. similar behaviour was found in real and hypothetical settings). One study (S76) found criterion validity only when data were weighted by respondents' certainty. In S102, hypothetical bias was no longer significant when only respondents who believed their answers could influence policy decisions were included in the analysis. However, S17 still found significant hypothetical bias after adjusting for consequentiality. The remaining two outcomes (both from S4) used an experimental market to value the environmental features of a market good (a detergent) and compared hypothetical DCE shares with the experimental market shares. The study found the same market shares one month after the goods were traded in the market, but different shares after four months.[8]

While these results suggest that DCEs are unlikely to predict respondents' behaviour in non-hypothetical situations, they must be interpreted with caution since the laboratory or controlled experiments with which DCEs were compared may themselves fail to predict behaviour outside the laboratory (Carlsson, 2010). The use of students in many of these tests (S14, S17, S47, S48 and S94) bears little resemblance to the diverse contexts in which DCE are used. For many non-market goods, a simulated market may not provide a true criterion measure of welfare impacts, that is actual behaviour may not be the "gold standard" against which DCE outcomes should be assessed, when the goal of the DCE is to estimate welfare impacts rather than predict market behaviour, or when there are no intentions to create real markets. Likewise, when the good is associated with non-utilitarian values (Lo and Spash, 2013; Kenter et al., 2015), "real" DCEs may not reflect full welfare effects.

[8] However, the difference after four months could be due to changes in market conditions that occurred during those four months. That is, the difference does not necessarily invalidate the DCE. Indeed, while it is difficult to separate instability of preferences over time (e.g. due to changing conditions) from unreliability of DCE in eliciting those preferences, the shorter the period between test and re-test, the less likely it seems that conditions and therefore preferences would have changed.

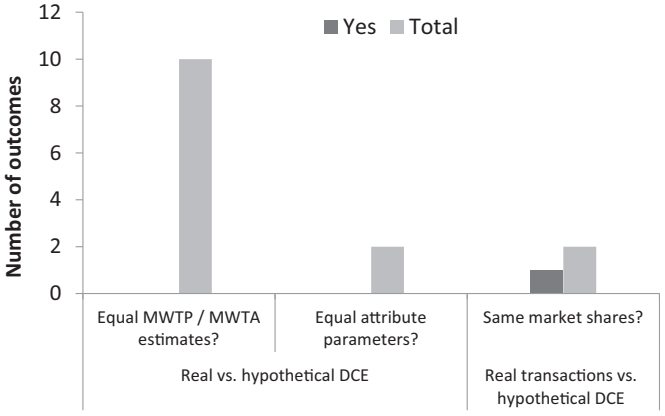


Fig. 4. Criterion validity tests: Do DCEs predict behaviour in real transactions? Note: Yes responses indicate that the outcomes are consistent with criterion validity (i.e. MWTP/MWTA estimates, attribute parameters or market shares were not significantly different between the hypothetical and real/simulated treatments at the 5% level).

4.2.2. Does DCE produce the same results as other methods?

Thirteen articles (producing 13 outcomes) tested for the convergent validity of DCE (S1, S6, S21, S26, S28, S33, S34, S67, S68, S73, S74, S87, S100), of which three were conducted in LICs (S73, S87, S100). The evidence generally supported consistency between DCE and other SP methods (Fig. 5). Two out of six outcomes comparing DCE and CVM did not find convergent compensating surplus estimates (S26 and S74). Equality of compensating surplus estimates depended on the specification of the utility function (S67) and the econometric modelling used (S28). Comparisons with other methods gave mixed outcomes: four contingent and qualitative ranking studies produced the same preference orderings as DCE (S6, S34, S68, S73); while multi-criteria analysis techniques produced different preference rankings than DCE (S68). MWTP estimates from DCE and hedonic pricing method were not shown to be statistically different (S87), whereas a significant difference was found in a comparison with the travel cost method (S1). While these results generally provide evidence of convergent validity between DCE measures and other SP approaches (CVM and contingent rankings), they only indicate 'validity by association', that is neither method can claim to be measuring the true value of the underlying construct (Bateman et al., 2002).

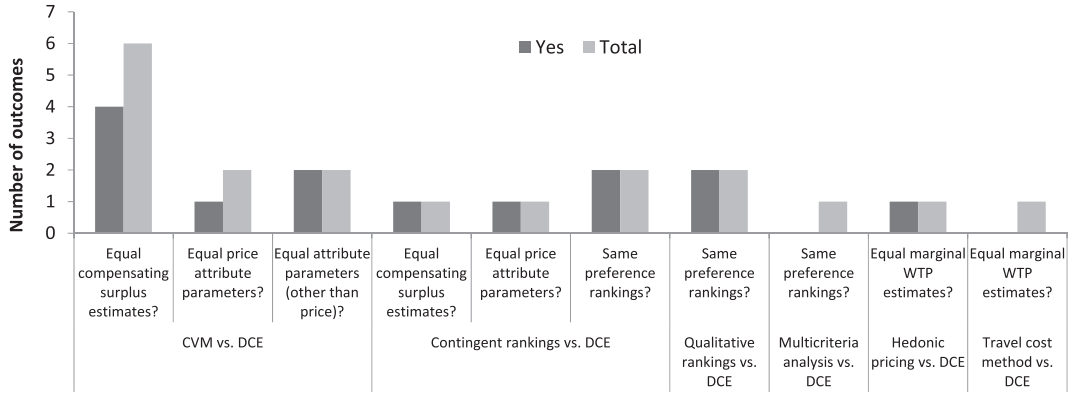


Fig. 5. Convergent validity tests: Does DCE produce the same results as other methods? Note: Yes responses indicate evidence consistent with convergent validity i.e. (MWTP/MWTA estimates – attribute parameters or market shares were not significantly different between the hypothetical and real/simulated treatments at 5% level[9]).

4.2.3. Do DCE results conform to theoretical expectations?

We found 30 articles that tested for the theoretical validity of DCE (of which three were conducted in LICs) producing 34 outcomes (supplement 5). Twenty-eight outcomes tested whether DCE results conformed to rational choice theory and six tested scope effects.

Twelve outcomes were quantitative measures of consistency with the continuity axiom (attribute attendance) as self-reported by respondents in follow-up statements (S2, S20, S24, S29, S30, S40, S53, S72, S84, S85, S97, S104). Between 15 and 100% of respondents reported that they did not behave as assumed by the axiom. Removing or accounting for discontinuous preferences in the analysis had mixed effects on WTP estimates; S24, S72, S84, and S97 found no systematic differences in MWTP between two models with and without consideration of ignored attributes whereas S20, S53, and S85 found the opposite. Accounting for stated partial attendance ('sometimes ignored') had statistically significant effects on both estimated preferences and welfare measures (S29). Non-attendance to alternatives due to unacceptable attribute levels occurred in 14% of the choices in S30. Design dimensions such as the number of choice sets, alternatives, levels and the level range did not affect stated attendance to attributes (S30 and S104). Two studies (S53 and S104) also employed econometric methods to reveal attribute non-attendance ex-post and found that a much smaller proportion of respondents was inferred to have attended to all attributes, compared to self-reported attendance. Two studies used qualitative techniques to test whether DCE results conform to the continuity axiom (S5 and S75). Post-DCE focus group results suggested that respondents attended to all the attribute levels across each of the alternatives, although some admitted that making trade-offs was difficult (S75). The verbal protocol and thinking aloud techniques showed that 66% of the sample followed compensatory rules, and greater emotional intensity significantly increased the likelihood of using non-compensatory decision making rules (S5).

Three studies reported measures of monotonicity, finding 25%, 21% and 1% of respondents violating that axiom respectively (S93, S86 and S65). The exclusion of choices that deviated from the monotonicity axiom resulted in reduced MWTP estimates in S86. Unstable preferences were diagnosed for 6–28% of respondents (S17, S18, S19, S84, S86, S93). Excluding respondents who violated the stability axiom from the analysis significantly lowered MWTP estimates in S86 but not in S84. S35, S95 and S96 found evidence of order effects, i.e. systematic changes in respondents' preference parameters related to the position of the choice task or the nature of options defined by attribute levels in previous tasks. While S32 found evidence of learning, attribute parameters and MWTP were not statistically different between choice tasks. Only one study assessed responses relative to the transitivity axiom and found that 17% of respondents showed intransitive preferences (S17). Six articles examined whether DCE conforms to expectations regarding scope effects. When examined across samples, four out of five DCE outcomes were not sensitive to scope (S43, S45, S61, and S70), while S58 found sensitivity to scope, except for the MWTP of one attribute for which test failure may be explained by its diminishing marginal values. In within-sample tests, MWTP of some attributes was sensitive to scope, while MWTP for others was not (S43, S70).

4.2.4. Do respondents protest about features of the surveys or find them incomprehensible or inconsequential?

In total 17 articles (producing 20 outcomes – see supplement 5) reported quantitative measures of the content validity of DCE, of which two were conducted in LICs (S7 and S12). Fifteen articles identified protest attitudes. 'Protesters' were defined as those who objected to the policy scenario in most studies (S7, S8, S77, S40, S44, S45, S46, S63, S64, S70, S95, S101, S103), those who perceived a lack of credibility of the hypothetical scenario (S8 and S90), and those who rejected the payment vehicle (S12). Respondents protesting ranged from 2% to 58% of the total sample in HICs and reached 90% in one of the LIC studies (S12). In one study (S7), respondents' average comprehension was rated at 3.1 (on a five point ascending scale) by enumerators, while in another (S90), 17% found making choices between alternative management options confusing and 40% of the respondents stated that they did not understand the valuation task. Three studies assessed the perceived consequentiality of the DCE survey and found that and 10%, 46%, and 62%, of the respondents believed that the study would not have an impact on policy (in and S17, S102 and S90, respectively). Only one article (S75) used qualitative techniques (post-DCE focus groups) to test for all three (protest attitudes, comprehensibility, and consequentiality). Participants in the focus group debriefings asserted that they considered their budget constraints and found the link between expected outcomes and the proposed policy realistic. However more than half of participants found the choice task difficult, with too much information.

4.3. Future directions in testing the reliability and validity of DCEs

The limited evidence base calls for greater attention to reliability and validity testing of DCEs in environmental valuation. We found that only 55% of the reliability outcomes passed the test. Reliability

[9] Here, we report compensating surplus estimates for CVM.

tests are essential to assess the robustness of results, and arguably as fundamental as calculating confidence intervals. Since DCE researchers may not be able to assess a priori how different information or designs will affect choices, tests of reliability should be incorporated into DCEs whenever resources allow. More specifically, we see research potential in the identification of minimum levels of deliberation for reliable preference elicitation in different contexts and more test-retests (see Fig. 3a), while acknowledging that care must be taken with regard to the effect on preferences of changes in economic conditions during the intervening period (see Footnote 8). Kenter et al. (2011), for example, conducted a DCE with rural and illiterate respondents and experimented with communal deliberative workshops to improve the quantity and quality of information available to participants. As between-sample tests currently dominate, more within-sample tests would strengthen the current evidence base.

Criterion validity is the least tested, yet often violated form of validity according to this review, with only 1 out of 13 outcomes passing the test. We recommend that whenever a reasonably valid and feasible criterion is available, DCE researchers should strive to measure hypothetical bias and investigate its sources. In other circumstances, methods should be developed to elicit value components that real markets and "real" DCE may not unveil, for example through participative and deliberative approaches (Spash, 2008) or mixed methods (Powe, 2007). Ultimately, for many non-market environmental goods no suitable criterion may ever become available. In such circumstances, SP techniques like hypothetical DCE may be the only option for monetary valuation, even if their criterion validity is untestable.

For convergent validity, 14 out of 19 outcomes passed the test, mostly when DCE is compared with other SP methods (CVM). Whilst there are many CVM versus revealed preference comparisons (Carson et al., 1996), only two studies compared DCE results with revealed preferences (hedonic pricing and travel cost method), with mixed findings (S1 and S87). We therefore concur with Lancsar and Swait (2014) recommendations for health economics: opportunities remain to compare revealed preference data with DCE estimates in environmental and resource economics, even though market failures undisputedly exist and revealed preference data cannot be presumed to provide a closer approximation to the "truth" than DCE data. Another avenue for further convergent validity testing, of which we found no existing study, would be to compare preferences revealed in response to interventions (e.g. randomized controlled trials where feasible) with those elicited by DCEs conducted prior to implementation. Ex-ante predictions from the original DCE could then be compared with ex-post revealed preference outcomes (Lancsar and Swait, 2014).

Theoretical validity tests, in particular attribute-attendance and, sensitivity-to-scope tests, are the most prevalent validity tests conducted to date, yet also often contested (Adamowicz et al., 2014). DCE analysis assumes behaviour compatible with rational choice theory, and deviations from rational choice theory have implications for analysis and interpretation. In particular, non-attendance to attributes has been a central issue in the examination of the theoretical validity of DCE as failure to identify and account for attribute non-attendance may bias welfare estimates and respondents' utility functions. Research from other fields suggests that respondents do not fully ignore attributes as they self-report but instead place lower importance on them, which need not be zero (Hess et al., 2013; Balcombe et al., 2014). Also, as insensitivity to scope has been extensively demonstrated for CVM (Carson, 2011) and was found in five out of six DCE studies in this review, we recommend that DCE researchers build in tests of how the environmental good is presented to the respondent whenever the results are suspected to be insensitive to scope.

Whether rational choice theory is a useful model of human behaviour has been much disputed by behavioural psychologists (e.g. Herrnstein, 1990) and economists equally assert that rational choice theory may not always correctly predict human behaviour. Recent advances in DCE modelling have suggested different ways to account for deviations from utility axioms (e.g. Campbell et al., 2011), however, they may not be a panacea if the level of non-conformities is unacceptably high. Once again, subjective judgments about what is acceptable must be made. Likewise, alternative choice theories or models which relax the assumptions of rational choice theory may be used (e.g. regret minimizing theory models, Thiene et al., 2011), however, they may pose problems for aggregation if the assumptions of the social welfare function used are violated. Ultimately, if DCEs are to be useful to policy and a lack of theoretical validity is a major concern, DCE researchers ought to gain a better understanding of the disparate and context-dependent ways in which respondents make choices (Loomes, 1999) as well as the factors or processes explaining violations of rational choice theory and how they relate to respondents' characteristics (Adamowicz et al., 2014). We recommend using qualitative approaches in combination with DCEs to make full use of key concepts in cognitive psychology and decision-making (Carlsson, 2010), for instance, to gain better understanding of choice processes and mechanisms (e.g. Clark et al., 2000; Powe et al., 2005). Qualitative approaches have so far been scarcely used to test the validity of DCE, because of concerns about the lack of generalizability and unclear economic interpretation of the results (Johnston, 2009) or possibly a lack of experience with qualitative approaches among DCE researchers. We believe that DCE and environmental valuation can benefit considerably from interdisciplinary approaches (Powe, 2007; Lancsar and Swait, 2014).

Evidence on the content validity of DCE is sparse with only 20 outcomes, which may be an artefact of our systematic review protocol, but could also imply a high level of undiagnosed protest beliefs and a need for more routine measurement. If a high number of respondents across DCE studies hold protest beliefs toward the payment vehicle or the policy scenario, this challenges the usefulness of the method in environmental decision-making. Similar concerns apply to perceived inconsequentiality and difficult-to-comprehend DCE survey designs, which may result in random responses instead of choices that would maximize utility. In particular, the identification of protesters is subjective and case study specific, and there is no agreement on how to handle protest attitudes in econometric modelling (Meyerhoff et al., 2014). We have observed a move towards ever more sophisticated econometric model specifications to analyse DCE data, but argue that survey design remains very important for improving DCE's reliability and validity. The use of debriefing questions is a simple but useful diagnostic tool to examine content validity, but we found them to be rarely reported. However, as with self-reported attribute-attendance, scholars have questioned the extent to which respondents' self-reported measures are reliable (Hess and Beharry-Borg, 2012).

Whilst the evidence is too heterogeneous to identify environmental goods for which reliability and validity are particularly problematic, we want to highlight the importance of testing reliability and validity in LICs for which we only found 12 studies. The scant evidence may be attributed to cost considerations, or a greater focus on delivering valuations commissioned for policy work rather than investigating methods (Whittington, 2010). At least until more evidence emerges, researchers should be particularly cautious when designing DCEs in LICs given the additional challenges that DCE researchers may face in these countries (Mangham et al., 2009).

Finally, we see deliberative methods as a promising approach to

understanding reliability and validity both in LICs and HICs. However, group-based deliberative approaches should be treated with caution since they may create scope for researcher-induced bias particularly when deliberation is used as part of the DCE. The "time-to-think" protocol (e.g. by Cook et al., 2007, a health economics' application) could avoid some of the drawbacks of participatory valuation and allow each individual to speak out and think free from wider group influence or social norms prevailing in group-based valuation approaches. Such "time-to-think" protocols could mimic reality better since respondents can talk to other household members and the survey setting is less restricting (Whittington, 2010).

4.4. Limitations of the systematic review approach

Although we took care to avoid missing relevant articles e.g. by using a test library, the search strings may be insufficiently sensitive to capture all available studies on the reliability and validity of DCE in the non-market environmental valuation literature. Adding more search terms might have permitted a more sensitive search, but would have been at the cost of specificity (Pullin and Stewart, 2006). The diverse ways in which reliability and validity are conceptualised and reported in the literature prevent a more comprehensive search without much greater resources. The use of consistent terminology in validity and reliability testing would assist future systematic reviews. Nevertheless, we believe that the results are representative of studies testing reliability and validity and provide a good assessment of the extent to which the peer-reviewed literature has reported empirical evidence of the reliability and validity of DCEs. Given the diversity and relative paucity of studies, especially the very small sub-samples for specific types of validity tests, we did not attempt a meta-analysis. Moreover, the very different contexts, treatments and DCE designs prohibit us from identifying factors that determine whether a specific method made DCEs more likely to be reliable and/or valid. Unless determining these factors was specifically the focus of a controlled test (within a study), such an analysis would need a large number of studies to control for confounding variables.

5. Conclusions

We systematically reviewed studies from 2003 to February 2016 that incorporated tests of reliability and validity of the DCE method when valuing non-market environmental goods. In these studies, DCE results were frequently susceptible to modest changes in survey designs and poorly predicted respondents' actual behaviour (albeit in somewhat artificial conditions). As expected, DCE outcomes were consistent with other SP based methods (mostly CVM) that share the same underlying theory. A considerable proportion of respondents' choices were inconsistent with the utility axioms assumed by DCEs, and evidence on the content validity of DCE was sparse. These results demonstrate a need to increase the evidence base on the reliability and validity of DCE in the environmental valuation literature. As DCE researchers always face uncertainties and difficulties in designing surveys, replicating reliability and validity tests would inform best practices in terms of alternative design approaches and give users of the DCE results, whether for policy-making or benefit transfer exercises, a sense of the level of confidence one can have in those results.

Despite the diverse, scant and inherently subjective nature of the evidence on the reliability and validity of DCE, it is sufficient to suggest considerable caution when using DCEs to inform decision-making. Arguably, the debate on the reliability and validity of DCE and other SP methods may never be settled as no decisive experiment exists (Whittington, 2010). Judgments about reliability and validity depend on not only the statistical significance of test results but also their economic importance. They are therefore specific to the context and intended uses of DCE, which are extremely diverse. In many environmental contexts, SP techniques may be the only valuation method available, and we expect that DCEs will continue to attract significant resources. However, their reliability and validity are still questionable and therefore require a similar level of attention. In particular, combining DCEs with revealed preference data is one promising research avenue in the environmental field that has been little explored. Likewise, the use of participative and deliberative processes, qualitative approaches, and other interdisciplinary techniques offer opportunities for improving the DCE method.

Acknowledgements

The authors would like to thank the financial support of the European Commission through the forest-for-nature-and-society (FONASO) joint doctoral programme. We also acknowledge the p4ges Project (NE/K010220/1) and ESPA Early Career Fellowship Grant (FELL-2014–104) both with support from the Ecosystem Services for Poverty Alleviation (ESPA) programme. The ESPA programme is funded by the Department for International Development (DFID), the Economic and Social Research Council (ESRC) and the Natural Environment Research Council (NERC). Thanks are also due to Jasper Kenter, Jette J. Jacobsen and an anonymous DCE researcher for commenting on the review protocol. We also thank Julia P. G. Jones and Jette J. Jacobsen for commenting on an early draft of this manuscript and gratefully acknowledge four anonymous reviewers for their helpful comments.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jenvman.2016.08.032.

References

Adamowicz, W., Glenk, K., Meyerhoff, J., 2014. Choice modelling research in environmental and resource economics. In: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar Publishing, Cheltenham, UK.

Adamowicz, W.L., 2004. What's it worth? an examination of historical trends and future directions in environmental valuation. Aust. J. Agric. Resour. Econ. 48 (3), 419–443. http://dx.doi.org/10.1111/j.1467-8489.2004.00258.x.

Alemu, M.H., Morkbak, M.R., Olsen, S.B., Jensen, C.L., 2013. Attending to the reasons for attribute non-attendance in choice experiments. Environ. Resour. Econ. 54 (3), 333–359. http://dx.doi.org/10.1007/s10640-012-9597-8.

Arana, J.E., Leon, C.J., 2009. Understanding the use of non-compensatory decision rules in discrete choice experiments: the role of emotions. Ecol. Econ. 68 (8–9), 2316–2326. http://dx.doi.org/10.1016/j.ecolecon.2009.03.003.

Azevedo, C., Corrigan, J.R., Crooker, J., 2009. Testing for the internal consistency of choice experiments using explicit rankings of quality attributes. In: Edelstein, A., Bär Hauppauge, D. (Eds.), Handbook of Environmental Research. Nova Science Publisher, New York, USA, pp. 507–517.

Balcombe, K., Bitzios, M., Fraser, I., Haddock-Fraser, J., 2014. Using attribute importance rankings within discrete choice experiments: an application to valuing bread attributes. J. Agric. Econ. 65 (2), 446–462. http://dx.doi.org/10.1111/1477-9552.12051.

Banerjee, S., Murphy, J.H., 2005. The scope test revisited. Appl. Econ. Lett. 12 (10), 613–617. http://dx.doi.org/10.1080/13504850500166253.

Barkmann, J., Glenk, K., Keil, A., Leemhuis, C., Dietrich, N., Gerold, G., Marggraf, R., 2008. Confronting unfamiliarity with ecosystem functions: the case for an ecosystem service approach to environmental valuation with stated preference methods. Ecol. Econ. 65 (1), 48–62. http://dx.doi.org/10.1016/j.ecolecon.2007.12.002.

Baskaran, R., Colombo, S., Cullen, R., 2013. Public preferences in irrigation and conservation development projects: does simultaneous consideration of substitutes in choice sets matter? Land Use Policy 33, 214–226. http://dx.doi.org/10.1016/j.landusepol.2013.01.004.

Bateman, I.J., Carson, R.T., Day, B., Hanemann, W.M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D.W., Sugden, R., Swanson, J., 2002. Economic Valuation with Stated Preferences Techniques: a Manual. Edward Elgar, Cheltenham, UK.

Bateman, I.J., Day, B.H., Jones, A.P., Jude, S., 2009. Reducing gain-loss asymmetry: a virtual reality choice experiment valuing land use change. J. Environ. Econ. Manag. 58 (1), 106–118. http://dx.doi.org/10.1016/j.jeem.2008.05.003.

Beharry-Borg, N., Hensher, D.A., Scarpa, R., 2009. An analytical framework for joint vs separate decisions by couples in choice experiments: the case of coastal water quality in tobago. Environ. Resour. Econ. 43 (1), 95–117. http://dx.doi.org/10.1007/s10640-009-9283-7.

Ben-Akiva, M., Morikawa, T., 1990. Estimation of travel demand models from multiple data sources. Transportation and traffic theory. In: Proceedings of the Eleventh International Symposium, Held July 18-20, 1990. Elsevier, Yokohama, Japan.

Bennett, J., Birol, E., 2010. Choice Experiments in Developing Countries: Implementation, Challenges and Policy Implications. Edward Elgar, Cheltenham, UK – Northampton, USA.

Bennett, J., Blamey, R., 2001. The Choice Modelling Approach to Environmental Valuation. Edward Elgar, Cheltenham, UK.

Birol, E., Koundouri, P., 2008. Choice Experiments Informing Environmental Policy. Edward Elgar, Cheltenham, UK, Northampton, MA, USA.

Boatman, N., Willis, K.G., Garrod, G., Powe, N., 2010. Estimating the Wildlife and Landscape Benefits of Environmental Stewardship: Final Report. The Food and Environment Research Agency, York, and the Centre for Research in Environmental Appraisal and Management, Newcastle University, Newcastle, Defra and Natural England.

Bradley, M., Daly, A., 1994. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation 21 (2), 167–184. http://dx.doi.org/10.1007/bf01098791.

Brewer, M., 2000. Research design and issues of validity. In: Reis, H., Judd, C. (Eds.), Handbook of Research Methods in Social and Personality Psychology. Cambridge University Press, Cambridge, pp. 11–26.

Bush, G., Colombo, S., Hanley, N., 2009. Should all choices count? using the cut-offs approach to edit responses in a choice experiment. Environ. Resour. Econ. 44 (3), 397–414. http://dx.doi.org/10.1007/s10640-009-9292-6.

Campbell, D., Hensher, D.A., Scarpa, R., 2011. Non-attendance to attributes in environmental choice analysis: a latent class specification. J. Environ. Plan. Manag. 54 (8), 1061–1076. http://dx.doi.org/10.1080/09640568.2010.549367.

Caparros, A., Oviedo, J.L., Campos, P., 2008. Would you choose your preferred option? Comparing choice and recoded ranking experiments. Am. J. Agric. Econ. 90 (3), 843–855. http://dx.doi.org/10.1111/j.1467-8276.2008.01137.x.

Carlsson, F., 2010. Design of stated preference surveys: is there more to learn from behavioral economics? Environ. Resour. Econ. 46 (2), 167–177. http://dx.doi.org/10.1007/s10640-010-9359-4.

Carlsson, F., Garcia, J.H., Lofgren, A., 2010. Conformity and the demand for environmental goods. Environ. Resour. Econ. 47 (3), 407–421. http://dx.doi.org/10.1007/s10640-010-9385-2.

Carson, R.T., 2011. Contingent Valuation: a Comprehensive Bibliography and History. Edward Elgar, Cheltenham, UK.

Carson, R.T., Czajkowski, M., 2014. The discrete choice experiment approach to environmental contingent valuation. In: Hess, S., Daly, A. (Eds.), Handbook of Choice Modelling. Edward Elgar, Northampton, pp. 202–235.

Carson, R.T., Flores, N.E., Martin, K.M., Wright, J.L., 1996. Contingent valuation and revealed preference methodologies: comparing the estimates for quasi-public goods. Land Econ. 72 (1), 80–99. DOI: 10.2307/3147159.

Carson, R.T., Groves, T., 2007. Incentive and informational properties of preference questions. Environ. Resour. Econ. 37 (1), 181–210. http://dx.doi.org/10.1007/s10640-007-9124-5.

Carson, R.T., Groves, T., 2011. Incentive and information properties of preference questions: commentary and extensions. In: Bennett, J. (Ed.), International Handbook of Non-market Environmental Valuation. Edward Elgar, Northampton, pp. 321–354.

Carson, R.T., Hanemann, W.M., 2005. Contingent valuation. In: Karl-Gran, M., Jeffrey, R.V. (Eds.), Handbook of Environmental Economics, 2. Elsevier, pp. 821–936.

Carson, R.T., Louviere, J.J., 2011. A common nomenclature for stated preference elicitation approaches. Environ. Resour. Econ. 49 (4), 539–559. http://dx.doi.org/10.1007/s10640-010-9450-x.

Christie, M., Azevedo, C.D., 2009. Testing the consistency between standard contingent valuation, repeated contingent valuation and choice experiments. J. Agric. Econ. 60 (1), 154–170. http://dx.doi.org/10.1111/j.1477-9552.2008.00178.x.

Christie, M., Fazey, I., Cooper, R., Hyde, T., Kenter, J.O., 2012. An evaluation of monetary and non-monetary techniques for assessing the importance of biodiversity and ecosystem services to people in countries with developing economies. Ecol. Econ. 83, 67–78. http://dx.doi.org/10.1016/j.ecolecon.2012.08.012.

Christie, M., Gibbons, J., 2011. The effect of individual 'ability to choose' (scale heterogeneity) on the valuation of environmental goods. Ecol. Econ. 70 (12), 2250–2257. http://dx.doi.org/10.1016/j.ecolecon.2011.07.011.

Christie, M., Hyde, T., Cooper, R., Fazey, I., Dennis, P., Warren, J., Gibbons, J., Hanley, N., 2010. An Economic Evaluation of the Ecosystem Service Benefits of the UK Biodiversity Action Plan, Report to Defra. Institute of Biological, Environmental and Rural Sciences. Aberystwyth University.

Clark, J., Burgess, J., Harrison, C.M., 2000. I struggled with this money business": respondents' perspectives on contingent valuation. Ecol. Econ. 33 (1), 45–62. http://dx.doi.org/10.1016/S0921-8009(99)00118-4.

Colombo, S., Christie, M., Hanley, N., 2013. What are the consequences of ignoring attributes in choice experiments? Implications for ecosystem service valuation. Ecol. Econ. 96, 25–35. http://dx.doi.org/10.1016/j.ecolecon.2013.08.016.

Cook, J., Whittington, D., Canh, D.G., Johnson, F.R., Nyamete, A., 2007. Reliability of stated preferences for cholera and typhoid vaccines with time to think in Hue, Vietnam. Econ. Inq. 45 (1), 100–114. http://dx.doi.org/10.1093/ei/cbl005.

Davidshofer, K.R., Murphy, Charles, O., 2005. Psychological of Testing: Principles and Applications. Pearson/Prentice Hall, Upper Saddle river.

Day, B., Bateman, I.J., Carson, R.T., Dupont, D., Louviere, J.J., Morimoto, S., Scarpa, R., Wang, P., 2012. Ordering effects and choice set awareness in repeat-response stated preference studies. J. Environ. Econ. Manag. 63 (1), 73–91. http://dx.doi.org/10.1016/j.jeem.2011.09.001.

Freeman, A.M., 2003. The Measurement of Environmental and Resource Values: Theory and Methods. Resources for the Future, Washington, DC.

Gracia, A., Loureiro, M.L., Nayga Jr., R.M., 2011. Are valuations from nonhypothetical choice experiments different from those of experimental auctions? Am. J. Agric. Econ. 93 (5), 1358–1373. http://dx.doi.org/10.1093/ajae/aar054.

Haddaway, N., Pullin, A., 2014. The policy role of systematic reviews: past, present and future. Springer Sci. Rev. 2 (1–2), 179–183. http://dx.doi.org/10.1007/s40362-014-0023-1.

Hanley, N., Barbier, E.M., 2009. Pricing Nature. Cost–benefit Analysis and Environmental Policy. Edward Elgar, Cheltenham.

Hanley, N., MacMillan, D., Wright, R.E., Bullock, C., Simpson, I., Parsisson, D., Crabtree, B., 1998. Contingent valuation versus choice experiments: estimating the benefits of environmentally sensitive areas in scotland. J. Agric. Econ. 49 (1), 1–15. http://dx.doi.org/10.1111/j.1477-9552.1998.tb01248.x.

Hanley, N., Mourato, S., Wright, R.E., 2001. Choice modelling approaches: a superior alternative for environmental valuatioin? J. Econ. Surv. 15 (3), 435–462. http://dx.doi.org/10.1111/1467-6419.00145.

Hausman, J., 2012. Contingent valuation: from dubious to hopeless. J. Econ. Perspect. 26 (4), 43–56. http://dx.doi.org/10.1257/jep.26.4.43.

Hensher, D.A., Rose, J.M., Greene, W., 2005. Applied Choice Analysis:A Primer. Cambridge university press, Cambridge.

Herrnstein, R.J., 1990. Rational choice theory: necessary but not sufficient. Am. Psychol. 45 (3), 356–367. http://dx.doi.org/10.1037/0003-066X.45.3.356.

Hess, S., Beharry-Borg, N., 2012. Accounting for latent attitudes in willingness-to-pay studies: the case of coastal water quality improvements in tobago. Environ. Resour. Econ. 52 (1), 109–131. http://dx.doi.org/10.1007/s10640-011-9522-6.

Hess, S., Daily, A. (Eds.), 2014. Handbook of Choice Modelling. Edward Elgar, Northampton, MA. DOI: 10.4337/9781781003152.

Hess, S., Hensher, D.A., Daly, A., 2012. Not bored yet – revisiting respondent fatigue in stated choice experiments. Transp. Res. Part A-Policy Pract. 46 (3), 626–644. http://dx.doi.org/10.1016/j.tra.2011.11.008.

Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., Caussade, S., 2013. It's not that I don't care, I just don't care very much: confounding between attribute non-attendance and taste heterogeneity. Transportation 40 (3), 583–607. http://dx.doi.org/10.1007/s11116-012-9438-1.

Johnston, R.J., 2009. Review of neil a. Powe: 'redesigning environmental valuation: mixing methods within stated preference techniques'. Ecol. Econ. 68 (5), 1564–1565. http://dx.doi.org/10.1016/j.ecolecon.2008.08.024.

Jones-Walters, L., Mulder, I., 2009. Valuing nature: the economics of biodiversity. J. Nat. Conservation 17 (4), 245–247. http://dx.doi.org/10.1016/j.jnc.2009.06.001.

Kenter, J.O., Hyde, T., Christie, M., Fazey, I., 2011. The importance of deliberation in valuing ecosystem services in developing countries—Evidence from the Solomon Islands. Glob. Environ. Change 21 (2), 505–521. http://dx.doi.org/10.1016/j.gloenvcha.2011.01.001.

Kenter, J.O., O'Brien, L., Hockley, N., Ravenscroft, N., Fazey, I., Irvine, K.N., Reed, M.S., Christie, M., Brady, E., Bryce, R., Church, A., Cooper, N., Davies, A., Evely, A., Everard, M., Fish, R., Fisher, J.A., Jobstvogt, N., Molloy, C., Orchard-Webb, J., Ranger, S., Ryan, M., Watson, V., Williams, S., 2015. What are shared and social values of ecosystems? Ecol. Econ. 111 (0), 86–99. http://dx.doi.org/10.1016/j.ecolecon.2015.01.006.

Kim, Y., Kling, C.L., Zhao, J., 2015. Understanding behavioral explanations of the wtp-wta divergence through a neoclassical lens: implications for environmental policy. Annu. Rev. Resour. Econ. 7 (1), 169–187. http://dx.doi.org/10.1146/annurev-resource-100913-012501.

Lancaster, K.J., 1966. A new approach to consumer theory. J. Political Econ. 74 (2), 132–157.

Lancsar, E., Swait, J., 2014. Reconceptualising the external validity of discrete choice experiments. PharmacoEconomics 32 (10), 951–965. http://dx.doi.org/10.1007/s40273-014-0181-7.

Laurans, Y., Rankovic, A., Billé, R., Pirard, R., Mermet, L., 2013. Use of ecosystem services economic valuation for decision making: questioning a literature blindspot. J. Environ. Manag. 119, 208–219. http://dx.doi.org/10.1016/j.jenvman.2013.01.008.

Le Gentil, E., Mongruel, R., 2015. A systematic review of socio-economic assessments in support of coastal zone management (1992–2011). J. Environ. Manag. 149, 85–96. http://dx.doi.org/10.1016/j.jenvman.2014.10.018.

Liebe, U., Meyerhoff, J., Hartje, V., 2012. Test-retest reliability of choice experiments in environmental valuation. Environ. Resour. Econ. 53 (3), 389–407. http://dx.doi.org/10.1007/s10640-012-9567-1.

Lo, A.Y., Spash, C.L., 2013. Deliberative Monetary Valuation: in search of a democratic and value plural approach to environmental policy. J. Econ. Surv. 27 (4), 768–789. http://dx.doi.org/10.1111/j.1467-6419.2011.00718.x.

Loomes, G., 1999. Some lessons from past experiments and some challenges for the future. Econ. J. 109 (453), F35–F45.
Louviere, J., Hensher, D.A., Swait, J.D., 2000. Stated Choice Methods: Analysis and Applications. Cambridge University Press, Cambridge.
Louviere, J.J., Flynn, T.N., Carson, R.T., 2010. Discrete choice experiments are not conjoint analysis. J. Choice Model. 3 (3), 57–72. http://dx.doi.org/10.1016/S1755-5345(13)70014-9.
Louviere, J.J., Pihlens, D., Carson, R., 2011. Design of discrete choice experiments: a discussion of issues that matter in future applied research. J. Choice Model. 4 (1), 1–8. http://dx.doi.org/10.1016/S1755-5345(13)70016-2.
Lusk, J.L., Schroeder, T.C., 2004. Are choice experiments incentive compatible? A test with quality differentiated beef steaks. Am. J. Agric. Econ. 86 (2), 467–482. http://dx.doi.org/10.1111/j.0092-5853.2004.00592.x.
Mahieu, P.A., Andersson, H., Beaumais, O., Crastes, R., Wolff, F.C., 2014. Is Choice Experiment Becoming More Popular than Contingent Valuation ? a Systematic Review in Agriculture, Environment and Health. FAERE Working Paper, 2014.12.
Mangham, L.J., Hanson, K., McPake, B., 2009. How to do (or not to do) Designing a discrete choice experiment for application in a low-income country. Health Policy Plan. 24 (2), 151–158. http://dx.doi.org/10.1093/heapol/czn047.
Mas-Colell, A., Whinston, M., Green, J., 1995. Microeconomic Theory. Oxford University Press, New York.
Meyerhoff, J., Liebe, U., 2009. Status quo effect in choice experiments: empirical evidence on attitudes and choice task complexity. Land Econ. 85 (3), 515–528.
Meyerhoff, J., Mørkbak, M., Olsen, S., 2014. A meta-study investigating the sources of protest behaviour in stated preference surveys. Environ. Resour. Econ. 58 (1), 35–57. http://dx.doi.org/10.1007/s10640-013-9688-1.
Michaud, C., Llerena, D., Joly, I., 2013. Willingness to pay for environmental attributes of non-food agricultural products: a real choice experiment. Eur. Rev. Agric. Econ. 40 (2), 313–329. http://dx.doi.org/10.1093/erae/jbs025.
Millenium Ecosystem Assessment, 2005. Ecosystems and Human Well-being: a Framework for Assessment. World Resources Institute, Washington, DC.
Mitchell, R.C., Carson, R.T., 1989. Using Surveys to Value Public Goods: the Contingent Valuation Method. Resources for the Future, Washington, DC.
Moran, D., McVittie, A., Allcroft, D.J., Elston, D.A., 2007. Quantifying public preferences for agri-environmental policy in Scotland: a comparison of methods. Ecol. Econ. 63 (1), 42–53. http://dx.doi.org/10.1016/j.ecolecon.2006.09.018.
Olsen, S.B., 2009. Choosing between internet and mail survey modes for choice experiment surveys considering non-market goods. Environ. Resour. Econ. 44 (4), 591–610. http://dx.doi.org/10.1007/s10640-009-9303-7.
Petrokofsky, G., Kanamaru, H., Achard, F., Goetz, S.J., Joosten, H., Holmgren, P., Lehtonen, A., Menton, M.C., Pullin, A.S., Wattenbach, M., 2012. Comparison of methods for measuring and assessing carbon stocks and carbon stock changes in terrestrial carbon pools. How do the accuracy and precision of current methods compare? A systematic review protocol. Environ. Evid. 1 (1), 1–21. http://dx.doi.org/10.1186/2047-2382-1-6.
Poe, G.L., Giraud, K.L., Loomis, J.B., 2005. Computational methods for measuring the difference of empirical distributions. Am. J. Agric. Econ. 87 (2), 353–365. http://dx.doi.org/10.1111/j.1467-8276.2005.00727.x.
Powe, N.A., 2007. Redesigning Environmental Valuation: Mixing Methods within Stated Preference Techniques. Edward Elgar, Cheltenham.
Powe, N.A., Garrod, G.D., McMahon, P.L., 2005. Mixing methods within stated preference environmental valuation: choice experiments and post-questionnaire qualitative analysis. Ecol. Econ. 52 (4), 513–526.
Pullin, A.S., Stewart, G.B., 2006. Guidelines for systematic review in conservation and environmental management. Conserv. Biol. 20 (6), 1647–1656. http://dx.doi.org/10.1111/j.1523-1739.2006.00485.x.
Ready, R.C., Champ, P.A., Lawton, J.L., 2010. Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. Land Econ. 86 (2), 363–381.
Robinson, J., Clouston, B., Suh, J., Chaloupka, M., 2008. Are citizens' juries a useful tool for assessing environmental value? Environ. Conserv. 35 (4), 351–360. http://dx.doi.org/10.1017/S0376892908005213.
Rolfe, J., Bennett, J., 2009. The impact of offering two versus three alternatives in choice modelling experiments. Ecol. Econ. 68 (4), 1140–1148. http://dx.doi.org/10.1016/j.ecolecon.2008.08.007.
Rolfe, J., Wang, X., 2011. Dealing with scale and scope issues in stated preference experiments. In: Benett, J. (Ed.), The International Handbook on Non-market Environmental Valuation. Edward Elgar, Cheltenham, U K, Northampton, MA, USA, pp. 254–272.
Scarpa, R., Ruto, E.S.K., Kristjanson, P., Radeny, M., Drucker, A.G., Rege, J.E.O., 2003. Valuing indigenous cattle breeds in Kenya: an empirical comparison of stated and revealed preference value estimates. Ecol. Econ. 45 (3), 409–426. http://dx.doi.org/10.1016/S0921800903000946.
Schaafsma, M., Brouwer, R., Liekens, I., De Nocker, L., 2014. Temporal stability of preferences and willingness to pay for natural areas in choice experiments: a test-retest. Resour. Energy Econ. 38, 243–260. http://dx.doi.org/10.1016/j.reseneeco.2014.09.001.
Schuman, H., Presser, S., 1981. Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Content. Academic Press, New York.
Smith, V.K., 2007. Judging quality. In: Kanninen, B. (Ed.), Valuing Environmental Amenities Using Stated Choice Studies, 8. Springer, Netherlands, pp. 297–333.
Solino, M., Farizo, B.A., Vazquez, M.X., Prada, A., 2012. Generating electricity with forest biomass: consistency and payment timeframe effects in choice experiments. Energy Policy 41, 798–806. http://dx.doi.org/10.1016/j.enpol.2011.11.048.
Spash, C.L., 2008. Deliberative monetary valuation and the evidence for a new value theory. Land Econ. 84 (3), 469–488.
Swait, J., Louviere, J., 1993. The role of the scale parameter in the estimation and comparison of multinomial logit-models. J. Mark. Res. 30 (3), 305–314. DOI: 10.2307/3172883.
Taylor, L.O., Morrison, M.D., Boyle, K.J., 2010. Exchange rules and the incentive compatibility of choice experiments. Environ. Resour. Econ. 47 (2), 197–220. http://dx.doi.org/10.1007/s10640-010-9371-8.
Tonsor, G.T., Shupp, R.S., 2011. Cheap talk scripts and online choice experiments: "looking beyond the mean. Am. J. Agric. Econ. 93 (4), 1015–1031. http://dx.doi.org/10.1093/ajae/aar036.
Torres, C., Hanley, N., Riera, A., 2011. How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. J. Environ. Econ. Manag. 62 (1), 111–121. http://dx.doi.org/10.1016/j.jeem.2010.11.007.
Vossler, C.A., Doyon, M., Rondeau, D., 2012. Truth in consequentiality: theory and field evidence on discrete choice experiments. Am. Econ. J. Microeconomics 4 (4), 145–171. http://dx.doi.org/10.1257/mic.4.4.145.
Whittington, D., 2010. What have we learned from 20 Years of stated preference research in less-developed countries? Annu. Rev. Resour. Econ. 2, 209–236. http://dx.doi.org/10.1146/annurev-resource-012809-103908.
Willis, K.G., Garrod, G., Scarpa, R., Powe, N., Lovett, A., Bateman, I.J., Hanley, N., Macmillan, D., 2003. The Social and Environmental Benefit of Forests in Great Britain. Forestry Commission, Edinburgh.